Categorial Grammar Induction with Stochastic Category Selection

Christian Clark, William Schuler
The Ohio State University
{clark.3664, schuler.77}@osu.edu

Abstract

Grammar induction, the task of learning a set of syntactic rules from minimally annotated training data, provides a means of exploring the longstanding question of whether humans rely on innate knowledge to acquire language. Of the various formalisms available for grammar induction, categorial grammars provide an appealing option due to their transparent interface between syntax and semantics. However, to obtain competitive results, previous categorial grammar inducers have relied on shortcuts such as part-of-speech annotations or an ad hoc bias term in the objective function to ensure desirable branching behavior. We present a categorial grammar inducer that eliminates both shortcuts: it learns from raw data, and does not rely on a biased objective function. This improvement is achieved through a novel stochastic process used to select the set of available syntactic categories. On a corpus of English child-directed speech, the model attains a recall-homogeneity of 0.48, a large improvement over previous categorial grammar inducers.

Keywords: Grammar induction, categorial grammar, language acquisition

1. Introduction

The learnability of linguistic structure by minimally supervised models is a question of enduring interest for both natural language processing and theoretical linguistics. Studies using recent Transformer-based (Vaswani et al., 2017) large language models (LLMs) suggest that these models are increasingly able to learn subtle syntactic phenomena such as subject-verb agreement and filler-gap dependencies (Linzen and Baroni, 2021; Wilcox et al., 2022)—casting doubt on influential claims (e.g., Chomsky, 1965) about the poverty of the stimulus and the importance of innate linguistic knowledge (Piantadosi, 2023). However, because LLMs do not produce grammars with clearly defined rules, it is difficult to characterize what kinds of structure they acquire. Furthermore, the extremely large scale of their training data makes LLMs unrealistic models of human language acquisition.

In contrast, grammar induction models provide a more explicit account of how linguistic structure is learned, with potentially greater relevance for understanding language acquisition. Although recent induction studies have often worked with probablistic context-free grammars (PCFGs; Kim et al., 2019; Zhu et al., 2020; Zhao and Titov, 2020), categorial grammars offer the advantage of a clean syntax–semantics interface, opening up possibilities for joint models of syntactic and semantic acquisition. But previous work on categorial grammar induction has faced other limitations. Earlier models relied on part-of-speech annotations (e.g., Bisk and Hockenmaier, 2012), unlike PCFG inducers that can learn from raw data. A recent study (Clark and Schuler, 2023) learns from raw data, but relies on an ad hoc bias term added to the objective function in order to achieve acceptable results.

We address these limitations by proposing a categorial grammar inducer that (1) learns from raw data, and (2) achieves competitive results without an ad hoc directional bias. This is accomplished by a novel method for selecting syntactic categories (Section 4). To evaluate the inducer as a model of child language acquisition, experiments are performed on corpora of child-directed speech in English (Section 5).

2. Related Work

Despite being considered a difficult task (Carroll and Charniak, 1992), grammar induction has been an active area of research for several decades (Lari and Young, 1990; Klein and Manning, 2002). Recent PCFG systems have achieved improvements through neural network architectures and other innovations such as multimodal grounding (Zhao and Titov, 2020; Zhang et al., 2021, 2022). Prior to Clark and Schuler (2023), earlier categorial grammar studies focused on learning from minimal POS annotations and/or a small number of labeled data points (Bisk and Hockenmaier, 2012, 2013; Bisk et al., 2015). Zettlemoyer and Collins (2005) present a system that learns a Combinatory Categorial Grammar (Steedman, 2000) to help with the task of mapping a sentence to its logical form—illustrating the potential for categorial grammars to serve as an aid in models of meaning acquisition, although this system requires explicit logical forms as input.

A related line of research has used formal methods to examine which classes of grammars are provably learnable under various assumptions about the input data and learner (e.g., Gold, 1967; Kanazawa, 1994; Clark and Yoshinaka, 2016). Kanazawa studies categorial grammars, showing that certain varieties such as rigid grammars are learnable from strings. The present study differs from this research in proposing a system that learns from broad-coverage, natural-

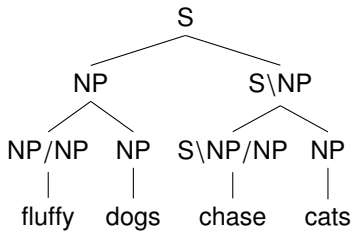


Figure 1: Example parse tree using a basic categorial grammar.

language corpora rather than idealized data.

3. Background: Induction Model

3.1. Grammar Formalism

The model presented in this paper uses a basic categorial grammar or AB-grammar (Ajdukiewicz, 1935; Bar-Hillel, 1953). This type of grammar includes a set of primitive categories, such as S or NP; two type-combining operators, \ and /; and two corresponding composition operations, backward function application and forward function application. Type-combining operators allow complex categories such as S\NP/NP to be formed from primitive categories.

Figure 1 shows an example parse tree from a basic categorial grammar with the primitive categories S and NP as well as complex categories NP/NP, S\NP, and S\NP/NP. Forward function application occurs when *fluffy* combines with *dogs* and when *chase* combines with *cats*; backward function application occurs when *fluffy dogs* combines with *chase cats*.

3.2. Objective Function

The induction model's objective function also follows previous work (Jin et al., 2021a; Clark and Schuler, 2023). It is defined as the marginal probability of the sentences in the dataset:

$$\mathrm{P}(\sigma) = \sum_{\tau,\tau'} \prod_{\eta\in\tau} \mathrm{P}(c_\eta \rightarrow c_{\eta 1}\ c_{\eta 2}) \cdot \prod_{\eta\in\tau'} \mathrm{P}(c_\eta \rightarrow w_\eta) \quad (1)$$

Here, σ is a single sentence. A possible parse tree for σ in Chomsky Normal Form can be divided into a set of nodes τ undergoing nonterminal expansions $c_\eta \rightarrow c_{\eta 1}\ c_{\eta 2}$ and a set of nodes τ' undergoing terminal expansions $c_\eta \rightarrow w_\eta$, where c_η is the nonterminal category at node η and w_η is the word located at node η.[1]

Bernoulli distributions determine whether a category c_η undergoes nonterminal or terminal expansion:

$$\mathrm{P}(\mathrm{Term} \mid c_\eta) = \underset{\{0,1\}}{\mathrm{softmax}}(\mathrm{N}_{\mathrm{Term}}(\mathbf{E}\,\delta_{c_\eta})) \quad (2)$$

[1]The variable $\eta \in \{1,2\}^*$ is a Gorn address specifying a path of left and right branches from the root node of the parse tree.

In this equation, c_η is a nonterminal category and δ_{c_η} is a vector representing a Kronecker delta function with 1 at index c_η and 0 elsewhere. $\mathbf{E} \in \mathbb{R}^{d\times|C|}$ is a matrix of nonterminal category embeddings of size d, where C is the set of nonterminal categories. $\mathrm{N}_{\mathrm{Term}}$ is a residual network containing 2 blocks (Kim et al., 2019).

Binary-branching nonterminal expansion probabilities are defined as follows:

$$\begin{aligned}\mathrm{P}(c_\eta \rightarrow c_{\eta 1}\ c_{\eta 2}) = \mathrm{P}(\mathrm{Term}{=}0 \mid c_\eta) \cdot \\ \mathrm{P}(c_\eta \rightarrow c_{\eta 1}\ c_{\eta 2} \mid c_\eta, \mathrm{Term}{=}0),\end{aligned} \quad (3)$$

with left- and right-child argument categories associated with weights $\mathbf{W}_\mathrm{L}, \mathbf{W}_\mathrm{R}$ and biases $\mathbf{b}_\mathrm{L}, \mathbf{b}_\mathrm{R}$:[2]

$$\begin{aligned}\mathrm{P}(c_\eta \rightarrow c_{\eta 1}\ c_{\eta 2} \mid c_\eta, \mathrm{Term}{=}0) = \\ \underset{(c',o)\in C_{\mathrm{arg}}\times\{\mathrm{L},\mathrm{R}\}}{\mathrm{softmax}}\left(\begin{bmatrix}\mathbf{W}_\mathrm{L}\\ \mathbf{W}_\mathrm{R}\end{bmatrix}\delta_{c_\eta} + \begin{bmatrix}\mathbf{b}_\mathrm{L}\\ \mathbf{b}_\mathrm{R}\end{bmatrix}\right)\end{aligned} \quad (4)$$

$C_{\mathrm{arg}} \subset C$ is the set of possible argument categories, and $o \in \{\mathrm{L},\mathrm{R}\}$ expresses the location of the argument child relative to the functor child.

Lexical unary-expansion rule probabilities are computed as follows:

$$\begin{aligned}\mathrm{P}(c_\eta \rightarrow w_\eta) = \mathrm{P}(\mathrm{Term}{=}1 \mid c_\eta) \cdot \\ \mathrm{P}(c_\eta \rightarrow w_\eta \mid c_\eta, \mathrm{Term}{=}1),\end{aligned} \quad (5)$$

with a softmax taken over words in the vocabulary:

$$\mathrm{P}(c_\eta \rightarrow w_\eta \mid c_\eta, \mathrm{Term}{=}1) = \underset{w_\eta}{\mathrm{softmax}}(\mathrm{N}'(\mathbf{E}\,\delta_{c_\eta})) \quad (6)$$

Here, N′ is residual network, similar to $\mathrm{N}_{\mathrm{Term}}$ except that the output layer's dimension is the size of the vocabulary.

4. Selection of Syntactic Categories

Along with defining an objective function, categorial grammar induction requires selecting an appropriate set of available syntactic categories. Because of constraints on how categories can combine, this decision has a potentially large effect on the final performance of the induction model.

Clark and Schuler (2023) select categories by setting a fixed number of primitives and maximum category depth.[3] The total number of available categories $|C_{P,D}|$ with P primitives and maximum depth D can be calculated from the following recurrence relation:

$$|C_{P,0}| = P \quad (7)$$

$$|C_{P,i}| = 2|C_{P,i-1}|^2 + P \quad (8)$$

[2]Note that the bias vectors are randomly initialized and thus do not enforce any particular branching behavior.

[3]Depth is defined according to a category's tree-based representation. For example, Figure 2(c) is a category of depth 2.

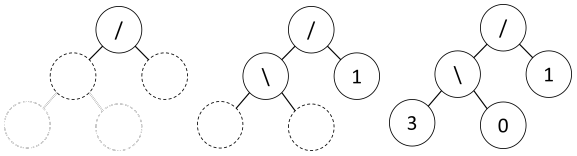


Figure 2: Steps to generate the complex category 3\0/1 following the stochastic generation process.

Clark and Schuler set P and D to 3 and 2, respectively, yielding 885 categories. The small number of primitives in their system is a potential disadvantage, considering that existing categorial grammars typically include more primitives. For example, the generalized categorial grammar from Nguyen et al. (2012) includes 14 primitive category types. CCGBank (Hockenmaier and Steedman, 2007) only uses 4 primitive category types, but these are paired with additional features to make finer distinctions, e.g. between declarative sentences and questions. In a basic categorial grammar, a larger set of primitives would be needed to capture these distinctions.

We therefore experiment with an alternative category selection method that allows for a larger number of primitives relative to the full number of categories. This method generates categories by a stochastic process based on two parameters p and q. Intuitively, p determines the preference for simple versus complex categories, while q determines the preference for a small or large set of primitives.

The process begins either by generating a primitive category n with probability $p(1-q)^n q$, or a function category with probability $1-p$ (the input and output categories of which will then recursively be generated).[4] The expression $p(1-q)^n q$ consists of a probability p of generating a primitive (as opposed to a function category), followed by an exponentially decreasing probability of generating a particular primitive with probability q after bypassing n previous primitives with probability $1-q$. The category set is then defined to be the set of categories whose probability according to this stochastic process is greater than some threshold.

This formulation has two desirable consequences. First, categories with less complexity are preferred, which is consistent with models such as the Prange et al. (2021) supertagger, as well as Bayesian inducers that use the infinite hidden Markov model (Beal et al., 2002). Second, categories using low-index primitives (e.g. 0 or 1) are assigned higher probabilities than categories using high-index primitives. This mirrors the fact that complex categories in handwritten grammars tend to reuse a small set of primitives (e.g. S\NP/NP/NP for ditransitive verbs).

[4]Primitive categories in the induction system are identified with integer labels $0, 1, 2, \ldots$.

Figure 2 illustrates the generation process for the example category 3\0/1. At step (a), a function category is generated with probability $1-p$. At (b), the left-hand child generates a second function category with probability $1-p$, and the right-hand child generates the primitive category 1 with probability $p(1-q)q$. Finally, at (c), the new function category's left child generates primitive category 3 with probability $p(1-q)^3 q$, and its right child generates primitive category 0 with probability pq. The overall probability of this category is $(1-p)^2 p^3 (1-q)^4 q^3$.

5. Experiments

5.1. Child-directed corpora

Following other recent grammar induction work (Jin et al., 2021a; Clark and Schuler, 2023), the induction model was tested on child-directed speech from the Adam and Eve sections (Brown, 1973) of CHILDES (MacWhinney, 2000). Both sections are in English. The Adam section was used for hyperparameter tuning; it contains a total of 28,780 sentences, with the child's age ranging from 2 years and 3 months to 5 years and 2 months. The Eve section was used for final testing; it comprises 14,251 sentences, with the child's age ranging from 1 year and 6 months to 2 years and 3 months. Syntactic annotations for these two sections of CHILDES were provided by Pearl and Sprouse (2013).

5.2. Procedures

The two evaluation metrics used were (1) unlabeled F1 score, which measures the alignment between predicted and annotated constituents; and (2) recall-homogeneity (Jin et al., 2021b), the product of unlabeled recall and homogeneity. Homogeneity, a metric used in part-of-speech tagging applications, measures to what degree a single induced category corresponds to a single annotated category.

The Adam corpus was used to perform a grid search to determine the optimal values for the p and q parameters described in Section 4, as well as a probability threshold. In general, lower values of q performed better, while the choice of p seemed to have less effect. The best-performing set of categories used $p = 0.5$, $q = 0.01$, and a probability threshold of approximately 8.2×10^{-6}, resulting in 2445 categories. See Appendix A for more information on the grid search.

The selected set of 2445 categories includes 638 primitives and 1807 single-operator complex categories. However, because only primitives 0 through 41 appear in complex categories, it is impossible for primitives 42 through 637 to appear in parses of multiword sentences.

This category set was subsequently used for testing on the Eve section of CHILDES. Appendix B

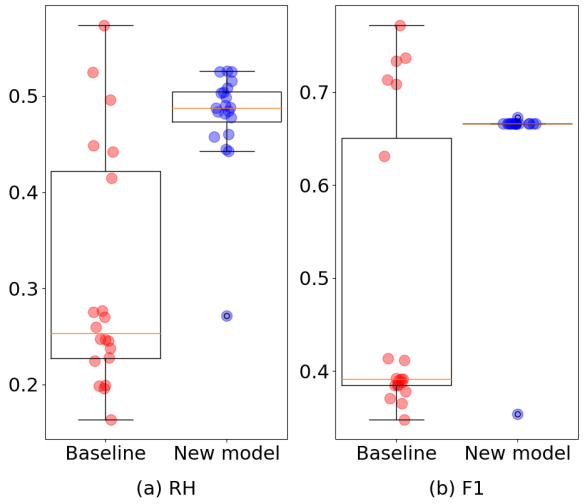


Figure 3: Recall-homogeneity (a) and F1 (b) on the Eve corpus, including 20 randomly initialized runs from the baseline and new models.

reports additional hyperparameters in the induction model.

5.3. Results

Figure 3 presents the main results, collected from 20 randomly initialized runs using the set of 2445 categories. These results were compared against 20 baseline runs using the Experiment 1 model from Clark and Schuler (2023) with 885 categories. Mean RH improves from the 0.31 in the baseline model to 0.48 in the new model, and mean F1 improves from 0.48 to 0.65. A permutation test showed that the improvements were significant ($p<0.01$). For comparsion, Jin et al. (2021a) report a mean RH of 0.49 from their PCFG inducer (F1 is not reported).

The baseline system shows a bimodal distribution in Figure 3, with a cluster of 6 runs averaging RH=0.48 and F1=0.72, and the remaining 14 runs averaging RH=0.23 and F1=0.39. In contrast, 19 out of 20 runs from the new system have comparable RH and F1 scores to the better runs from the baseline system.

The one remaining run from the new system is a clear outlier, with RH=0.27 and F1=0.35. This run uses backward function application and left-branching structures far more often than the other 19 runs, which mostly rely on forward function application and right branching; see Figure 4 for an example sentence. The pattern of better runs using right branching and worse runs using left branching is similar to what occurs with the baseline model. However, it is encouraging that only a single run from the new system tends to produce left-branching trees; it suggests that the new method of category selection is more effective at guiding the model toward desirable right-branching analyses. Notably, this is accomplished

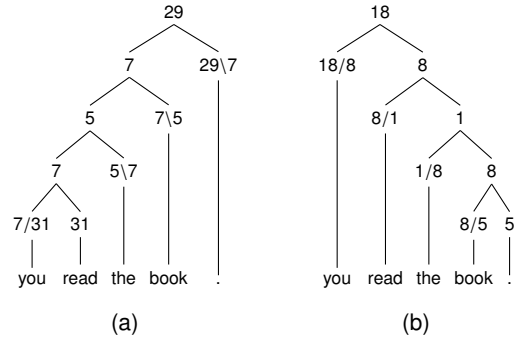


Figure 4: Comparison of predicted parses for an example sentence in Eve from the lowest-RH ((a); RH=0.27) and highest-RH ((b); RH=0.53) runs of the grammar inducer.

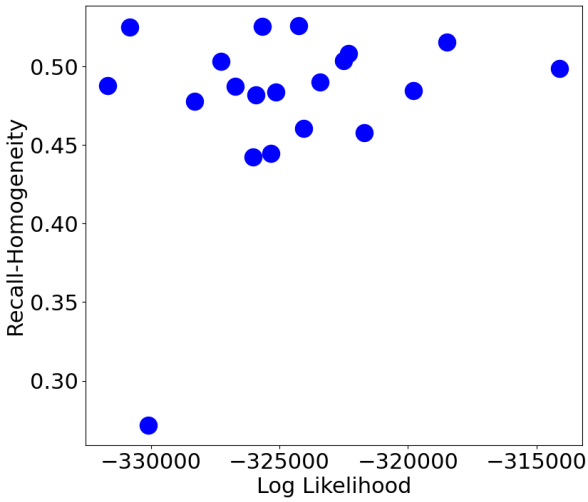


Figure 5: Relationship between log likelihood and recall-homogeneity across the 20 runs of the new model on Eve.

without any hard-coded directional bias term like Clark and Schuler (2023) use in their Experiment 2.

The relationship between log likelihood and RH is illustrated in Figure 5. The run with RH=0.27 has a relatively low log likelihood relative to other runs. However, the 19 other runs do not show much of a correlation between the two measures—suggesting further room for improvement in defining an objective function that steers the inducer toward accurate parses.

Inspection of the most frequent induced categories reveals that each is associated with a small set of annotated categories, most often just one (Table 1). The table shows that all of the most frequent categories are primitives. This occurs because the selected set of 2445 categories—comprising only primitives and single-operator complex categories—only allows complex categories to appear at preterminal nodes in parse trees (as happens in Figure 4). It can also be observed that low-index primitive categories are used for a range of annotated categories such as NP and

Induced category	Annotated categories (relative freq.)
1	NP (0.22)
8	VP (0.55)
0	VP (0.30), S (0.19), SQ (0.08)
7	PP (0.29)
4	S (0.09)
26	ROOT (0.98)
30	ROOT (1.00)
18	ROOT (1.00)
10	SBAR (0.16), S (0.1)
31	ROOT (1.00)

Table 1: Annotated categories from the Penn Treebank tag set associated with each of the 10 most frequent induced categories (sorted from most to least frequent). Induced categories come from the best-performing induction model (RH=0.53). Annotated categories that are associated with an induced category at least 5% of the time are reported.

VP, while high-index primitives are almost exclusively used for ROOT. This likely reflects the fact that high-index primitives appear in fewer complex categories overall, thanks to their lower probability according to the stochastic process. Appendix C presents confusion matrices relating the most frequent induced and annotated syntactic categories.

6. Conclusion

We test a categorial grammar induction model that uses a novel technique for category selection. On a corpus of child-directed speech, this model attains an average RH of 0.48, a large improvement over the Clark and Schuler (2023) system that brings the model's performance to the level of state-of-the art PCFG inducers. Predictions from the model lend support to the idea that syntactic structure may be learnable without extensive prior knowledge, and show interesting correlations between induced and annotated categories.

Ethics Statement

We do not anticipate any ethical concerns from this work.

Acknowledgments

We thank the anonymous reviewers for their feedback. This work was supported by the National Science Foundation grant #2313140. All views expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation.

7. Bibliographical References

Kazimierz Ajdukiewicz. 1935. Die syntaktische konnexitat. In S. McCall, editor, *Polish Logic 1920-1939*, pages 207–231. Oxford University Press. Translated from Studia Philosophica 1: 1–27.

Yehoshua Bar-Hillel. 1953. A quasi-arithmetical notation for syntactic description. *Language*, 29:47–58.

Matthew J. Beal, Zoubin Ghahramani, and Carl E. Rasmussen. 2002. The infinite hidden markov model. In *Machine Learning*, pages 29–245. MIT Press.

Yonatan Bisk, Christos Christodoulopoulos, and Julia Hockenmaier. 2015. Labeled grammar induction with minimal supervision. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 870–876.

Yonatan Bisk and Julia Hockenmaier. 2012. Simple robust grammar induction with combinatory categorial grammars. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Yonatan Bisk and Julia Hockenmaier. 2013. An HDP model for inducing combinatory categorial grammars. *Transactions of the Association for Computational Linguistics*, 1:75–88.

R. Brown. 1973. *A First Language*. Harvard University Press, Cambridge, MA.

Glenn Carroll and Eugene Charniak. 1992. Two Experiments on Learning Probabilistic Dependency Grammars from Corpora. *Working Notes of the Workshop on Statistically-Based {NLP} Techniques*, pages 1–13.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, Mass.

Alexander Clark and Ryo Yoshinaka. 2016. Distributional learning of context-free and multiple context-free grammars. *Topics in grammatical inference*, pages 143–172.

Christian Clark and William Schuler. 2023. Categorial grammar induction from raw data. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2368–2379, Toronto, Canada. Association for Computational Linguistics.

E Mark Gold. 1967. Language identification in the limit. *Information and Control*, 10(5):447–474.

Julia Hockenmaier and Mark Steedman. 2007. CCGbank: a corpus of CCG derivations and dependency structures extracted from the Penn Treebank. *Computational Linguistics*, 33(3):355–396.

Lifeng Jin, Byung-Doh Oh, and William Schuler. 2021a. Character-based PCFG induction for modeling the syntactic acquisition of morphologically rich languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4367–4378, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lifeng Jin, Lane Schwartz, Finale Doshi-Velez, Timothy Miller, and William Schuler. 2021b. Depth-Bounded Statistical PCFG Induction as a Model of Human Grammar Acquisition. *Computational Linguistics*, 47(1):181–216.

Makoto Kanazawa. 1994. *Learnable classes of categorial grammars*. Stanford University.

Yoon Kim, Chris Dyer, and Alexander Rush. 2019. Compound probabilistic context-free grammars for grammar induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2369–2385.

Dan Klein and Christopher D. Manning. 2002. A generative constituent-context model for improved grammar induction. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Karim Lari and Steve J Young. 1990. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer speech & language*, 4(1):35–56.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7(1):195–212.

Brian MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk*, third edition. Lawrence Elrbaum Associates, Mahwah, NJ.

Luan Nguyen, Marten van Schijndel, and William Schuler. 2012. Accurate unbounded dependency recovery using generalized categorial grammars. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING '12)*, pages 2125–2140, Mumbai, India.

Lisa Pearl and Jon Sprouse. 2013. Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*, 20:23–68.

Steven Piantadosi. 2023. Modern language models refute Chomsky's approach to language. *Lingbuzz Preprint, lingbuzz*, 7180.

Jakob Prange, Nathan Schneider, and Vivek Srikumar. 2021. Supertagging the Long Tail with Tree-Structured Decoding of Complex Categories. *Transactions of the Association for Computational Linguistics*, 9:243–260.

Mark Steedman. 2000. *The syntactic process*. MIT Press/Bradford Books, Cambridge, MA.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2022. Learning syntactic structures from string input. In *Algebraic Structures in Natural Language*, pages 113–138. CRC Press.

Luke Zettlemoyer and Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of the Proceedings of the Twenty-First Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-05)*, pages 658–666, Arlington, Virginia. AUAI Press.

Songyang Zhang, Linfeng Song, Lifeng Jin, Haitao Mi, Kun Xu, Dong Yu, and Jiebo Luo. 2022. Learning a grammar inducer from massive uncurated instructional videos. *arXiv preprint arXiv:2210.12309*.

Songyang Zhang, Linfeng Song, Lifeng Jin, Kun Xu, Dong Yu, and Jiebo Luo. 2021. Video-aided unsupervised grammar induction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1513–1524, Online. Association for Computational Linguistics.

Yanpeng Zhao and Ivan Titov. 2020. Visually grounded compound PCFGs. *arXiv preprint arXiv:2009.12404*.

Hao Zhu, Yonatan Bisk, and Graham Neubig. 2020. The return of lexical dependencies: Neural lexicalized PCFGs. *Transactions of the Association for Computational Linguistics*, 8:647–661.

A. Grid search for category selection

To select p and q values for the stochastic category selection process detailed in Section 4, a grid search was performed on the Adam corpus using $p, q \in \{0.001, 0.01, 0.1, 0.2, 0.5, 0.8, 0.9, 0.99, 0.999\}$. Because testing all pairs of p and q would have been too costly, the range of p values was tested with q fixed at 0.5, and then the range of q values was tested with p fixed at 0.5. For each tested pair of p and q, probability thresholds were set to select roughly 100, 1000, or 2500 categories. The probability threshold was the minimal value t such that the number of categories with probability greater than or equal to t was no more than 100, 1000, or 2500.

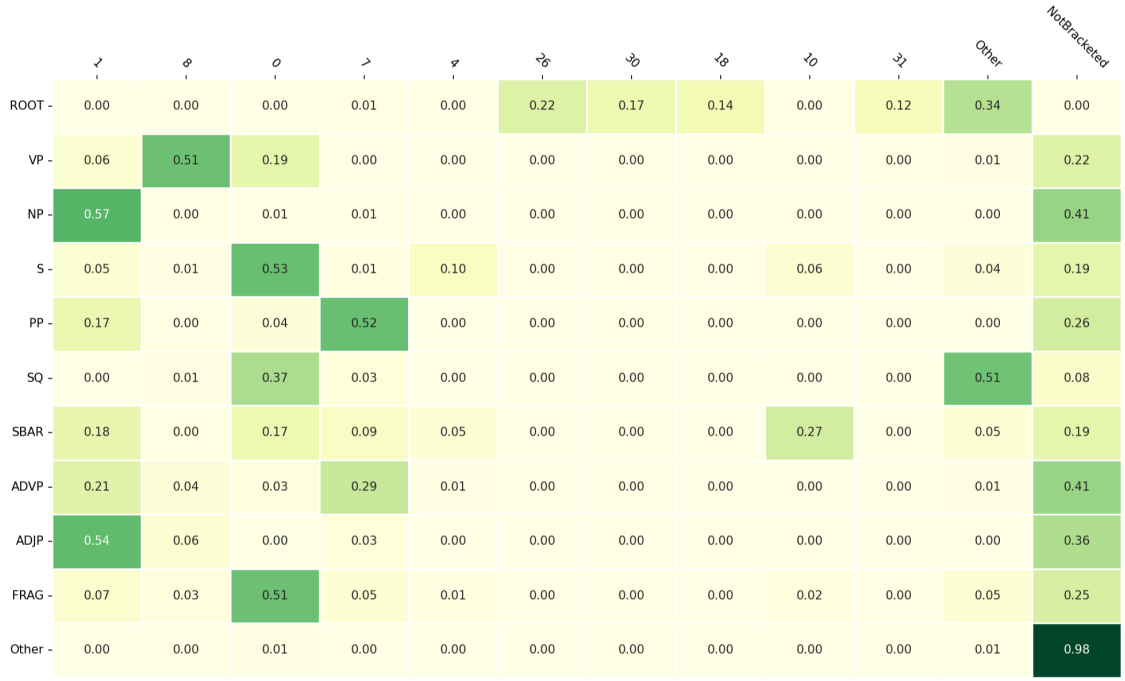
After the initial grid search, lower probability thresholds permitting up to 7500 categories were tested with p and q set to 0.5 and 0.01. However, the set of 2445 categories performed best on Adam. To avoid underflow, probabilities were log-transformed during category selection.

B. Hyperparameters

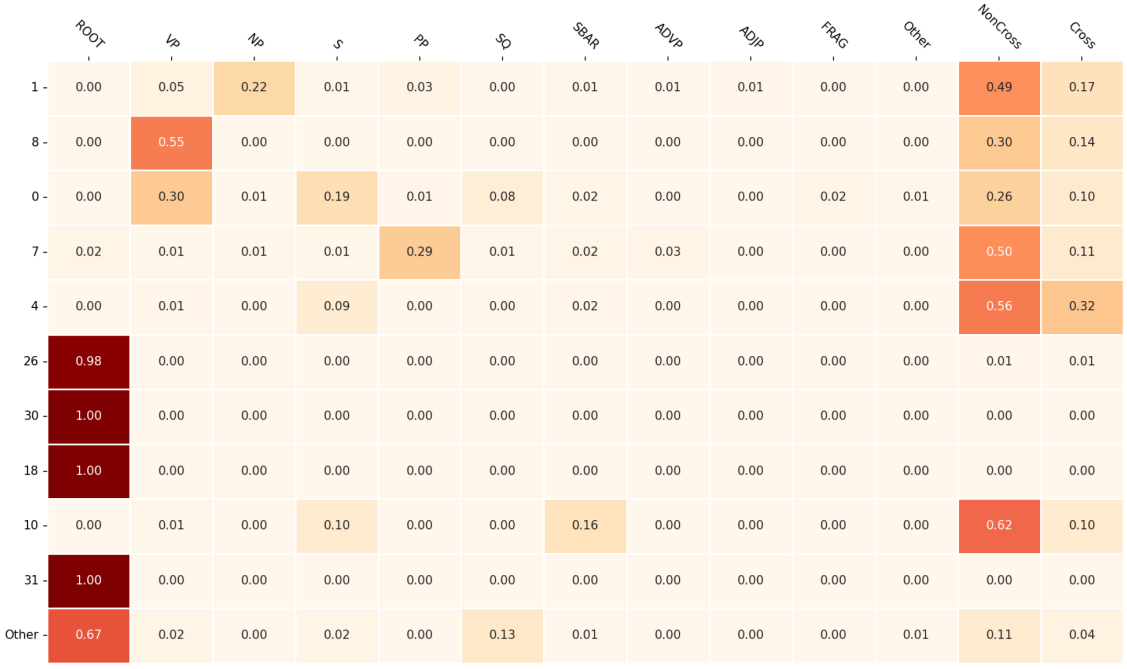
Hyperparameters in the induction model matched those reported in Clark and Schuler (2023). The Adam optimizer was used with a learning rate of 0.0001. The category embedding size and hidden state size were set to 64. Each run was randomly initialized and run for 20 epochs with a batch size of 2 sentences.

C. Comparison of Induced and Annotated Categories

Figure 6 on the following page presents confusion matrices comparing the most frequent induced and annotated syntactic categories.



(a) Recall



(b) Precision

Figure 6: Comparison of frequent induced and annotated categories in the Eve corpus. Induced categories came from the best-performing induction model with 2445 categories (RH=0.53). “NotBracketed” in (a) refers to phrases of a particular category that were not bracketed together in the predicted parse. “NonCross” in (b) counts phrases belonging to an induced category that did not appear as constituent in the annotated parse, but did not cross constituent boundaries in the annotation. “Cross” counts phrases that did cross annotated constituent boundaries.